Chatbot Abuse Typologies: A Scoping Review

Cassandra Kist
University of Strathclyde
Glasgow, United Kingdom
Cassandra.kist@strath.ac.uk

Emma Nicol
University of Strathclyde
Glasgow, United Kingdom
Emma.nicol@strath.ac.uk

Catherine Chavula
University of Strathclyde
Glasgow, United Kingdom
Catherine.chavula@strath.ac.uk

Abstract

Chatbots can increase service efficiency and reduce operational costs in many industries, such as healthcare, banking, retail, as well as in tourism and leisure. However, the increase in chatbot use creates additional opportunities for chatbot abuse. Abuse directed at chatbots is a complex phenomenon entangled with ethical implications, potentially resulting in harm to humans/users. Despite a growing body of related research, there are no purpose-made frameworks or typologies for describing and analyzing chatbot abuse. Prompted by the absence of such a theoretical tool, we conducted a scoping review to synthesize previous research and begin building a typology of chatbot abuse, identifying how chatbot abuse and associated harms have been described and labelled, and the topics addressed by chatbot abuse research. Our work also identifies additional factors (the relational nature of chatbot abuse and the marginalized role of interaction contexts) that may enrich such a typology, suggesting future research trajectories.

CCS Concepts

• **Human-centered computing → HCI theory, concepts and models**; **Natural language interfaces**; **Interaction paradigms**.

Keywords

Chatbot abuse, chatbot harms, typology of harms, chatbot ethics

1 Introduction

Despite the widespread use of chatbots across different industries and a growing body of related literature, there is a lack of understanding regarding chatbot abuse, that is, harm directed towards chatbots. This includes how chatbot abuse should be defined and described and what behaviors, motivations, and contexts are associated with its occurrence [21]. In particular, while previous research on user engagement with chatbots in heritage contexts highlighted examples of agent abuse [43], there has been no subsequent development of a shared understanding and explanation in the broader literature. The abuse of chatbots raises concerns about the possible harmful effects on human-human relationships, as seen when robots are abused [50]. In turn, it prompts questions about user motivations and how understanding users, chatbot design and context can inform strategies to mitigate or prevent related harms.

The legal, ethical, and moral questions raised by the abuse of chatbots have been a concern for researchers since the early 2000s [18, 19]. More recently, as technology has evolved, the focus of research in related areas such as human-robot interaction has been to try to classify abusive behaviors and harms and to develop frameworks for describing and analyzing abuse [42, 50]. Such frameworks and typologies are useful for supporting shared understandings of behaviors and harms, comparing empirical and conceptual types, predicting possibilities, and developing theoretical explanations. Prompted by the lack of shared understanding of chatbot abuse between researchers, developers, and deployers of chatbots, we undertook a scoping literature review to understand and synthesize how chatbot abuse has been described in the literature. Specifically, we addressed the following research questions:

RQ1: How has chatbot abuse been described and defined in previous research?
RQ2: What topics are being addressed by research in/on/related to chatbot abuse?
RQ3: What are the current gaps and overlaps in how chatbot abuse has been defined/described across this research?

2 Study Design

We chose a scoping review to synthesize evidence about chatbot abuse, how it has been defined and described, and the topics addressed by the related literature. Scoping reviews can be useful when 'future research planning requires a clear picture of the availability and gaps in existing research' [36], and for defining a research concept and the boundaries of a research topic [2]. We searched Scopus, IEEE, Web of Science and ACM for literature detailing harms towards chatbots, specifically through the lens of abuse. We used the search terms: 'Abuse AND chatbots'; 'Abuse AND conversational agents'. In the case of ACM, a large number of irrelevant results were returned (>1000), so we narrowed our search criteria to only include literature with the search terms in the abstract, resulting in 229 papers being found from the database searches. Our inclusion criteria were broad, one key criterion being that papers had to explicitly reference abuse or harm directed towards a chatbot or conversational agent such as a robot. Table 1 provides an overview of our search results and study corpus.

To ensure the relevance of papers to our research questions, we applied the inclusion criteria in two steps. First, titles and abstracts of these 229 papers were screened. Where relevance was unclear from the title and abstract, we conducted full-text reviews to assess eligibility. The authors resolved ambiguities through discussion, resulting in 27 eligible papers from the database search (Table 1). We then performed manual searches to find relevant papers that may not have included our search terms in titles or abstracts [2]. We used backward chaining (examining reference lists of eligible papers) and forward chaining (using Google Scholar to check citations), finding an additional 20 papers. This combination of database and manual searches identified 47 papers that discussed abuse directed

This work is licensed under a Creative Commons CC0 1.0 Universal License.
Mensch und Computer 2025 – Workshopband, Gesellschaft für Informatik e.V., 31. August - 03. September 2025, Chemnitz, Germany
© 2025 Copyright held by the owner/author(s). Publication rights licensed to GI.
https://doi.org/10.18420/muc2025-mci-wip-307

Table 1: Literature collected from search protocol

Database	Total articles screened	Relevant	Literature citation
Scopus	89	17	[1, 5, 7, 8, 10, 11, 15–18, 20, 21, 24, 25, 27, 31, 46]
Web of Science	59	1	[34]
IEEE	72	8	[6, 12, 14, 23, 26, 30, 43, 45]
ACM	9 (abstract search only)	1	[9]
Manual search		20	[4, 19, 22, 28, 29, 32, 33, 35, 37–43, 47–51]

towards conversational agents. Using our research questions as an analytical frame for reviewing the literature, we took notes regarding RQ1: how chatbot abuse is described and defined and RQ2: topics addressed by the literature, based on the papers' main aim(s), thematically coding our literature notes. For RQ3, we zoomed in on types of chatbot abuse, collating similar types/descriptions of harms, and examining differences to identify overlaps and gaps across the literature. To do this, we collated descriptions of chatbot abuse from the literature into a framework intended to categorize and label online harm [3] (Table 2). The typology of online harms by Banko et al. [3] usefully provides a broad range of categories to define and categorize harmful behaviors that aligned with the descriptions of chatbot abuse found in the literature.

3 Results and Discussion

3.1 RQ1: Definitions of Chatbot Abuse

Previous research studies have described and categorized different forms of chatbot abuse, either defining abuse for lab-based experiments or, as the main interest of the research. For instance, in studies aimed at mitigating chatbot abuse through design, definitions/categorizations of verbal abuse that may be directed towards chatbots were created and classified as either, insult, swear, or threat [9] and in another, as using offensive language, represented through a list of offensive terms [35]. In contrast, some studies have attempted to categorize and describe observed user behaviors through the analysis of interactions [28, 33, 43] and playful speculation [42]. The types of abusive behavior and harms identified, and the overlaps and potential gaps are explored in the following sections.

3.1.1 'Types' of Chatbot Abuse. Studies, such as [20, 48, 50] examining chatbot abuse, have either directly or indirectly described abuse and its associated harms. Across these studies there are overlaps in descriptions of abuse, but differences in terminology. For instance, as mentioned above and reflected in Table 2, much research has focused on what could be grouped as 'insults' [3], with researchers inconsistently labeling harms as insults, swears, flaming, threats, sexual abuse, and discriminatory opinions (related to gender/race) (Table 2). Other authors have described the testing of chatbots or conversational agents' abilities as abuse [43]. Researchers have also observed physical attacks on hardware [39], which is more evident in the abuse of robots but may extend to other conversational agents [18].

Moreover, several harms directed at chatbots that could be considered 'misuse' have been framed by some authors as chatbot 'abuse', particularly in research stemming from the cybersecurity field. For instance, [12], [44], and [45], suggest that chatbot abuse also includes bypassing fraudulent detection, security systems, and or ethics protocols to generate malicious content, and using the chatbot for malicious intent. Similarly, Derner et al. [23] describe harm directed toward chatbots as encompassing attacks on Large Language Models (LLMs) that would 'disrupt' the model or manipulate it to create unintended outputs such as through prompt injections/data poisoning etc. These differing perspectives on what constitutes harm towards chatbots reflect a shared ambiguity across this research area regarding how abuse should be described or labelled — for instance, as a swear or insult or more generally as either 'abuse' or 'misuse'.

Recent research has examined abuse in chatbot interactions and has represented associated 'types' through a more relational lens, examining how harm emerges during interactions. Zhang et al. [50] examined a dataset of conversations with the Replika generative AI chatbot app and identified 6 categories of 'algorithmic' harms mainly directed towards users. Within these harms, chatbots and users can play different roles in the emergence of harm, based on their initiation and involvement. Through a speculative workshop by Rezzani et al. [42], student participants identified four types of harm: outbursts, clashes, repressions, and oppression, that highlight how harm emerges in interactions, with user and robot taking on different roles. This work on relational harm prompts consideration regarding whether similar forms of harm are directed towards chatbots. While the attitudes of the students in [42] provide some sense of the 'harm' directed towards robots, broader user perspectives on what abuse of chatbots is, are also largely absent from this literature.

3.2 RQ2: Topics Addressed by Studies of Chatbot Abuse

Topics addressed by the literature encompass concerns for ethics, user motivations for chatbot abuse, ways to mitigate chatbot abuse and studies that categorize different forms of chatbot abuse.

3.2.1 Ethics. A concern for ethics is apparent in the literature collected: This moves beyond a concern simply for ethical treatment of humans by robots or chatbots, to contemplate concern for ethical treatment of these forms of artificial intelligence and the associated knock-on effects [4, 13, 17, 25]. Discrimination and abuse of robots for instance, is suggested to result in emotional harms to witnesses — including controllers/moderators of conversational agents [40], harm to society-at-large [25], potentially encouraging discrimination [17] and harm between humans [40], and also animals [25] and negative impacts on perceptions of chatbots' abilities [8]. Goo et al. [26] further consider how abuse of a conversational agent may have differential impacts on the distress of users/bystanders, with

females and participants who previously experienced victimization or sexism as experiencing more distress.

3.2.2 User Motivations. A further key topic examined by the literature is motivators for abuse, having both an ethical dimension and implications for chatbot design. Studies can largely be grouped into those that investigate predictors of chatbot abuse such as chatbot characteristics [5, 49], user characteristics [11], and user perceptions of chatbots [20, 40]. Several studies examine the effects of chatbot characteristics such as gender roles and gender appearance [46], physical/visual attractiveness [32], and ethnicity [38] on abusive interactions between users and chatbots. These studies reveal how discrimination that occurs between humans is similarly apparent in interactions between chatbots and humans, manifesting more in interactions where chatbots are feminized or portray a non-white ethnicity [38, 46]. Similarly, studies indicate that the level of anthropomorphism may increase instances of abuse, potentially related to anxiety caused by the uncanny valley effect and users' perceptions of chatbots [7]. Studies examining user characteristics have identified key predictors of chatbot abuse such as frequent use of chatbots [41], having studied a computer-related subject at university (correlated with lower experiences of guilt resulting from chatbot abuse) [11], gender, with males more likely to abuse chatbots [11], and ethical orientation e.g., an ideology of idealism was found to decrease the likelihood of chatbot abuse [41]. Studies have also revealed how users' perceptions of chatbot/robots' identities as, variously, human-like entities vs. property [40, 41]; a subject vs. object [42]; and a master vs. slave [20] impacts their abuse. Additionally, a few studies have theorized the nature of chatbot abuse as proactive or reactive [42], examining the role of chatbots in triggering abuse including communication breakdowns and service failures [51]. The context of chatbot engagement, in terms of its anonymity and privacy, has also been theorized as contributing to chatbot abuse due to feelings of disinhibition [17]. Surprisingly, despite some studies framing their research as focussing on social/relationship chatbots [11, 50] or highlighting the private and anonymous context [17] – few studies consider the interaction context of users nor consider differences across industrial or cultural contexts.

3.2.3 Chatbot Abuse Mitigation Strategies. A further recurrent topic in the literature reviewed is the design of chatbots to prevent or mitigate chatbot abuse. Here, the literature addresses characteristics of chatbots in terms of gender representation, level of anthropomorphism and communication strategy [7, 22, 35]. Li et al. [35] explored communication responses to offensive users and found that chatbots that respond empathetically and redirect the conversation significantly lower the re-offense rate of users when compared to the use of avoidance or counter responses. Another suggested mitigation strategy is making chatbots less anthropomorphic and their gender more androgynous [7]. A recent study showed how a robot's reaction to abuse - in this case shutting down - could motivate bystanders to intervene [47].

3.3 RQ3: Overlaps and Gaps in Typology

Here, we outline how our scoping review on chatbot abuse relates to existing typologies for online harm with a view to identifying overlaps and gaps in terminology. Table 2 collates terms and descriptions of chatbot abuse and maps these into subcategories and categories of harmful online behaviors as defined by Banko et al. [3]. Banko et al. [3] developed four broad categories of harm broken into subcategories, with consideration for the type of target (society, a person, or subject) and the way abuse is articulated (e.g., aggression, insult etc.). These categories encompass: hate and harassment, self inflicted harm, ideological harm, and exploitation. Of these four broad categories, 'Hate and Harassment' aligned most with descriptions of chatbot abuse in the literature as: directed at an individual or group and intended to cause the target harm (demean or undermine etc.) [3]. For example, in the literature collected, very similar descriptions of using inappropriate or offensive language are labelled as vulgarities [48], insults [9], and flaming [9], which is captured by Banko et al.'s [3] subcategory 'Insult'. Other harms directed at chatbots, described in similar ways but labelled differently are also reflected in the subcategories: Threat of Violence (e.g., hazing/testing), and Sexual Aggression (e.g., sexual harassment, sexual advances). Table 2, therefore, demonstrates how harm directed at chatbots has been described in similar ways but labelled inconsistently. Table 2 also reveals the attention that has been directed towards certain subcategories of abuse: for instance, 'insult' has the most references compared to other harm subcategories. Despite a few overlaps between Table 2 and abuse categories identified by other literature, it is unclear whether harms encompassed for example, by Zhang et al.'s [50] categories of algorithmic harm such as relational transgression are applicable to the abuse of chatbots.

4 Implications, Limitations and Future Work

Our research begins to build a typology to create a shared terminology for describing chatbot abuse — synthesizing research in this area and associated harms (Table 2). This table is used to demonstrate the overlaps and differences between how chatbot abuse is described and labeled in the literature. We demonstrate that while much research focuses on 'insults', additional harms have received less attention due to a lack of analysis of relational or emerging dimensions of chatbot abuse and the examination of interaction contexts. By reviewing the most common topics addressed and analyzed in this area, we identify an ethical concern for the (differential) impact of abuse on users/bystanders and society, a variety of motivations/pre-determinants for chatbot abuse, and mitigation strategies. We believe that a typology of chatbot abuse would enable researchers to further analyze, identify gaps, and cross-compare studies of chatbot abuse using a shared understanding and language for this phenomenon. In turn, this would enable designers and deployers of chatbots to identify relevant risks/mitigation strategies in design and deployment. As a scoping literature review, we have gathered insights into how chatbot abuse has been described and labelled, and the topics addressed in this area, to initiate the development of a typology and identify future research trajectories. Initial searches were limited to Scopus, IEEE, Web of Science, and ACM using predefined search terms and inclusion criteria and relied on the researchers' subjective judgments. This reflects a potential limitation in capturing the spectrum of chatbot abuse terminology. Additional insights could emerge from systematically reviewing

Table 2: Harms directed towards chatbots identified across the literature mapped to categories and subcategories of online harms created by Banko et al. [3]

Harm Category	Potential harm subcategory	Observed forms of harm in the scoping literature review	Description of observed harms
Hate and Harassment	Insult	Verbal abuse [20]	Encompasses sexual harassment and injuries
		Offensive/profane language [35]; Profanity [29]	Encompasses a wide range of potentially offensive words
		Insult [9]	Encompasses 'treating the agent with disrespect' (p. 2)
		Swearing [9]	Encompasses 'using offensive language' (p. 2)
		Flaming [43]	Described as 'direct insults, swearing and offensive utterances' (p. 1130)
		Vulgarities [48]	Expletives directed at the agent
	Threat of Violence	Threat [9]	Encompasses 'An intention to harm' (p. 2)
		Testing [28, 43]	Testing the system's boundaries through a factual or perceptual test
		Hazing [43]	Testing in such a way that it is more hostile/toying with a system
		Manipulating for malicious intent [12, 23, 44, 45]	Bypassing ethics, security or fraudulent detection system to create malicious content or for malicious intent
		Disrupting the model [23]	Affecting the outputs or reliability of the model through for example, data poisoning, intensive prompts
	Sexual Aggression	Sexual harassment [5, 20]	Terms related to sexuality, pornography and explicit sexual attention
		Sexual advances [31]	Sexual comment
		Sexually explicit language [48]	Sexual comments directed at the chatbot
	Identity attack	Gender identity [1, 5, 7]	Questions around issues of gender; comments associated with stereotyping
		Ethnic group [41]	Attacks oriented towards an ethnic group/identity

changes over time, charting usage of methods, research contexts, and evaluating how research was undertaken.

5 Conclusion

This paper provides a snapshot of research on chatbot abuse via a scoping review, including how chatbot abuse has been described and labeled and the key topics being addressed, and begins to synthesize this literature to develop a typology of chatbot abuse. In doing so, we contribute to the field of conversational user interfaces (CUI) and to broader human computer interaction (HCI) research by initiating the development of a shared typology and delineating what may be considered 'abuse' vs. 'misuse'. Finally, we identify future research trajectories that may inform understanding of chatbot abuse through attention to its relational nature and the significant role of interaction contexts. Chatbot abuse/abuse of conversational agents adversely affects witnesses and bystanders and can propagate abusive behaviors between humans, with negative impacts on human-human relationships. As chatbots become more prevalent, an enhanced understanding of chatbot abuse and a shared typology for grounding and guiding future research in this area is paramount.

References

[1] Antonella De Angeli and Sheryl Brahnam. 2008. I hate you! Disinhibition with virtual partners. *Interacting with Computers* 20, 3 (2008), 302–310. https://doi.org/10.1016/j.intcom.2008.02.004 Special Issue: On the Abuse and Misuse of Social Agents.

[2] Hilary Arksey, , and Lisa O'Malley. 2005. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* 8, 1 (Feb. 2005), 19–32. https://doi.org/10.1080/1364557032000119616 Publisher: Routledge _eprint: https://doi.org/10.1080/1364557032000119616.

[3] Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A Unified Taxonomy of Harmful Content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem (Eds.). Association for Computational Linguistics, Online, 125–137. https://doi.org/10.18653/v1/2020.alw-1.16

[4] Jessica K. Barfield. 2023. Discrimination against robots: Discussing the ethics of social interactions and who is harmed. *Paladyn* 14, 1 (2023), 1–12. https://doi.org/10.1515/pjbr-2022-0113

[5] Sheryl Brahnam and Antonella De Angeli. 2012. Gender affordances of conversational agents. *Interacting with Computers* 24, 3 (May 2012), 139–153. https://doi.org/10.1016/j.intcom.2012.05.001

[6] Sheryl Brahnam, Marianthe Karanikas, and Margaret Weaver. 2011. (Un)dressing the interface: Exposing the foundational HCI metaphor "computer is woman". *Interacting with Computers* 23, 5 (Sept. 2011), 401–412. https://doi.org/10.1016/j.intcom.2011.03.008

[7] Sheryl. Brahnam and Margaret Weaver. 2015. Re/framing virtual conversational partners: A feminist critique and tentative move towards a new design paradigm. In *Design, User Experience, and Usability: Users and Interactions*, Vol. 9187. Springer International Publishing, Cham, 172–183. https://doi.org/10.1007/978-3-319-20898-5_17

[8] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376461

[9] Hyojin Chin and Mun Yong Yi. 2019. Should an agent be ignoring it? A study of verbal abuse types and conversational agents' response styles. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312826

[10] HyoJin Chin and Mun Yong Yi. 2022. Voices that Care Differently: Understanding the Effectiveness of a Conversational Agent with an Alternative Empathy Orientation and Emotional Expressivity in Mitigating Verbal Abuse. *International Journal of Human–Computer Interaction* 38, 12 (2022), 1153–1167. https://doi.org/10.1080/10447318.2021.1987680 arXiv:https://doi.org/10.1080/10447318.2021.1987680

[11] Hyojin Chin and MMun Yong Yong Yi. 2024. Exploring the influence of user characteristics on verbal aggression towards social chatbots. *Behaviour and Information Technology* 44, 8 (2024), 1576–1594. https://doi.org/10.1080/0144929X.2024.2362957

[12] Minhaz Chowdhury, Nafiz Rifat, Shadman Latif, Mostofa Ahsan, Md Saifur Rahman, and Rahul Gomes. 2023. ChatGPT: The Curious Case of Attack Vectors' Supply Chain Management Improvement. In *2023 IEEE International Conference on Electro Information Technology (eIT)*. IEEE, Romeoville, IL, USA, 499–504. https://doi.org/10.1109/eIT57321.2023.10187385 ISSN: 2154-0373.

[13] Simon Coghlan, Frank Vetere, Jenny Waycott, and Barbara Neves. 2019. Could Social Robots Make Us Kinder or Crueller to Humans and Animals? *International Journal of Social Robotics* 11, 5 (2019), 741–751. https://doi.org/10.1007/s12369-019-00583-2

[14] Evan N. Crothers, Nathalie Japkowicz, and Herna L. Viktor. 2023. Machine-Generated Text: A Comprehensive Survey of Threat Models and Detection Methods. *IEEE Access* 11 (2023), 70977–71002. https://doi.org/10.1109/ACCESS.2023.3294090

[15] Amanada Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational AI. In *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*. Association for Computational Linguistics, Dominican Republic, 7388–7403.

[16] Amanda Curry and Verena Rieser. 2019. A crowd-based evaluation of abuse response strategies in conversational agents. In *SIGDIAL 2019 - 20th Annual Meeting of the Special Interest Group Discourse Dialogue - Proceedings of the Conference*. Association for Computational Linguistics (ACL), Sweden, Stockholm, 361–366. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091562953&partnerID=40&md5=524a0612b064d1a32f48141bac68033e

[17] Antonella De Angeli. 2009. Ethical implications of verbal disinhibition with CAs. *PsychNology Journal* 7 (Jan. 2009), 49–57.

[18] Antonella De Angeli and S. Brahnam. 2006. Sex stereotypes and conversational agents. In *Proceedings the Workshop on Gender & Interaction*. ACM, New York, NY, USA, 3. https://iris.unitn.it/handle/11572/90254?mode=full Accepted: 2015-04-21T17:39:18Z.

[19] Antonella De Angeli, Sheryl Brahnam, Peter Wallis, and Alan Dix. 2006. Misuse and abuse of interactive technologies. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, New York, NY, USA, 1647–1650. https://doi.org/10.1145/1125451.1125753

[20] Antonella De Angeli and R Carpenter. 2005. Stupid computer! Abuse and social identities. In *Proceedings of the INTERACT 2005 workshop "Abuse: The darker side of Human-Computer Interaction*. INTERACT, Manchester, UK, 19–25. https://www.academia.edu/16805067/Stupid_computer_Abuse_and_social_identities

[21] Roberta De Cicco. 2024. Exploring the Dark Corners of Human-Chatbot Interactions: A Literature Review on Conversational Agent Abuse. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 14524 LNCS. Springer Nature, Switzerland, 185–203. https://doi.org/10.1007/978-3-031-54975-5_11

[22] Luara De Grazia, Alex Peiró-Lilja, Mireia Farrús, and Mariona T. Delor. 2024. How should Conversational Agent systems respond to sexual harassment?. In *TEICAI 2024 - 1st Workshop Towards Ethical and Inclusive Conversational AI: Language Attitudes, Linguistic Diversity, and Language Rights, Proceedings of the Workshop*. Association for Computational Linguistics, TX, USA, 28–35. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85188661046&partnerID=40&md5=0a36c496bbd81a37de806e74eb6015a7

[23] Erik Derner, Kristina Batistič, Jan Zahálka, and Robert Babuška. 2024. A Security Risk Taxonomy for Prompt-Based Interaction With Large Language Models. *IEEE Access* 12 (2024), 126176–126187. https://doi.org/10.1109/ACCESS.2024.3450388

[24] Joachim Diederich. 2021. Risks to Artificial Intelligence. *Cognitive Systems Monographs* 42 (2021), 115–125. https://doi.org/10.1007/978-3-030-71842-8_7

[25] P. Goldie. 2010. The Moral Risks of Risky Technologies. *International Library of Ethics, Law and Technology* 5 (2010), 127–138. https://doi.org/10.1007/978-90-481-8647-1_8

[26] Hideki G. Goo, Katie Winkle, Tom Williams, and Megan Strait. 2023. Victims and Observers: How Gender, Victimization Experience, and Biases Shape Perceptions of Robot Abuse. In *IEEE International Workshop on Robot and Human Communication, RO-MAN*. IEEE, Busan, Korea, 2439–2446. https://doi.org/10.1109/RO-MAN57019.2023.10309640

[27] Agneta Gulz, Magnus Haake, Annika Silvervarg, Björn Sjödén, and George Veletsianos. 2010. Building a Social Conversational Pedagogical Agent-Design Challenges and Methodological Approaches. In *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices*. IGI Global, Pennsylvania, USA, 128–155.

[28] Joakim Gustafson and Linda Bell. 2000. Speech technology on trial: Experiences from the August system. *Natural Language Engineering* 6, 3-4 (Sept. 2000), 273–286. https://doi.org/10.1017/S1351324900002485

[29] Jennifer Hill, W. Randolph Ford, and Ingrid G. Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* 49 (Aug. 2015), 245–250. https://doi.org/10.1016/j.chb.2015.02.026

[30] Chunhui Hu and Jianfeng Chen. 2023. A Dimensional Perspective Analysis on the Cybersecurity Risks and Opportunities of ChatGPT-Like Information Systems. In *2023 International Conference on Networking and Network Applications (NaNA)*. IEEE, New York, USA, 324–331. https://doi.org/10.1109/NaNA60121.2023.00061

[31] Merel Keijsers, Christoph Bartneck, and Friederike Eyssel. 2021. What's to Bullying a Bot?: Correlates Between Chatbot Humanlikeness and Abuse. *Interaction Studies* 22, 1 (2021), 55–80. https://doi.org/10.1075/is.20002.kei

[32] Rabia Khan and Antonella De Angeli. 2009. The Attractiveness Stereotype in the Evaluation of Embodied Conversational Agents. In *Human-Computer Interaction – INTERACT 2009*, Tom Gross, Jan Gulliksen, Paula Kotzé, Lars Oestreicher, Philippe Palanque, Raquel Oliveira Prates, and Marco Winckler (Eds.). Springer, Berlin, Heidelberg, 85–97. https://doi.org/10.1007/978-3-642-03655-2_10

[33] Stefan Kopp, Lars Gesellensetter, Nicole Krämer, and Ipke Wachsmuth. 2005. A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application. In *Intelligent Virtual Agents*, Vol. 3661. Springer Nature, Berlin, Heidelberg, 329–343. https://doi.org/10.1007/11550617_28

[34] Jeehyun Jenny Lee. 2024. "Who is sexually harassed? A python code haha": imaginaries of a post-violent AI world. *Feminist Media Studies* 0, 0 (2024), 1–16. https://doi.org/10.1080/14680777.2024.2418393 Publisher: Routledge _eprint: https://doi.org/10.1080/14680777.2024.2418393.

[35] Haojun Li, Dilara Soylu, and Christpher Manning. 2021. Large-Scale Quantitative Evaluation of Dialogue Agents' Response Strategies against Offensive Users. In *SIGDIAL 2021 - 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*. Association for Computational Linguistics, Singapore and Online, 556–561.

[36] Craig Lockwood, Kelli Borgess dos Santos, and Robin Pap. 2019. Practical Guidance for Knowledge Synthesis: Scoping Review Methods. *Asian Nursing Research* 13, 5 (Dec. 2019), 287–294. https://doi.org/10.1016/j.anr.2019.11.002

[37] Xiaojuan Ma, Emily Yang, and Pascale Fung. 2019. Exploring Perceived Emotional Intelligence of Personality-Driven Virtual Agents in Handling User Challenges. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1222–1233. https://doi.org/10.1145/3308558.3313400

[38] Clifford Nass, Katherine . Isbister, and Eun-Ju Lee. 2001. Truth is beauty: researching embodied conversational agents. In *Embodied conversational agents*. MIT Press, Cambridge, MA, USA, 374–402.

[39] Tatsuya Nomura, Takayuki Kanda, Hiroyoshi Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. 2016. Why do children abuse robots? *Interaction Studies* 17, 3 (Jan. 2016), 347–369. https://doi.org/10.1075/is.17.3.02nom Publisher: John Benjamins.

[40] Jo Ann Oravec. 2023. Rage against robots: Emotional and motivational dimensions of anti-robot attacks, robot sabotage, and robot bullying. *Technological Forecasting and Social Change* 189 (April 2023), 122249. https://doi.org/10.1016/j.techfore.2022.122249

[41] Namkee Park, Kyungeun Jang, Seonggyeol Cho, and Jinyoung Choi. 2021. Use of offensive language in human-artificial intelligence chatbot interaction: The effects of ethical ideology, social competence, and perceived humanlikeness. *Computers in Human Behavior* 121 (Aug. 2021), 106795. https://doi.org/10.1016/j.chb.2021.106795

[42] Andrea Rezzani, Antonella De Angeli, Maria Menéndez Blanco, and Max Dorfmann. 2023. The Space of User Aggression in Human-Robot Interaction: A Theoretically Grounded Workshop on Imaginaries of Humans Against Robots. In *ACM International Conference Proceeding Series*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3605390.3605402

[43] Susan Robinson, D. Traum, Midhun Ittycheriah, and Joe Henderer. 2008. What would you Ask a conversational Agent? Observations of Human-Agent Dialogues in a Museum Setting. In *International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Marrakech, Morocco, 1125–1131. https://www.semanticscholar.org/paper/What-would-you-Ask-a-conversational-Agent-of-in-a-Robinson-Traum/28a07e4379ab54a2d779735c69776f58ce638576

[44] Sayak Saha Roy, Poojitha Thota, Krishna Vamsi Naragam, and Shirin Nilizadeh. 2024. From Chatbots to Phishbots?: Phishing Scam Generation in Commercial

Large Language Models. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, San Fransisco, CA, USA, 36–54. https://doi.org/10.1109/SP54263.2024.00182 ISSN: 2375-1207.

[45] Ashfak Md Shibli, Mir Mehedi A. Pritom, and Maanak Gupta. 2024. AbuseGPT: Abuse of Generative AI ChatBots to Create Smishing Campaigns. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*. IEEE, San Antonio, TX, USA, 1–6. https://doi.org/10.1109/ISDFS60797.2024.10527300 ISSN: 2768-1831.

[46] Annika Silvervarg, Kristin Raukola, Magnus Haake, and Agneta Gulz. 2012. The effect of visual gender on abuse in conversation with ECAs. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7502 LNAI (2012), 153–160. https://doi.org/10.1007/978-3-642-33197-8_16

[47] Xiang Zhi Tan, Marynel Vázquez, Elizabeth J. Carter, Cecilia G. Morales, and Aaron Steinfeld. 2018. Inducing Bystander Interventions During Robot Abuse with Social Mechanisms. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL, USA) *(HRI '18)*. Association for Computing Machinery, New York, NY, USA, 169–177. https://doi.org/10.1145/3171221.3171247

[48] George Veletsianos, Cassandra Scharber, and Aaron Doering. 2008. When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with Computers* 20, 3 (May 2008), 292–301. https://doi.org/10.1016/j.intcom.2008.02.007

[49] Catherine Zanbaka, Paula Goolkasian, and Larry Hodges. 2006. Can a virtual cat persuade you? the role of gender and realism in speaker persuasiveness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. Association for Computing Machinery, New York, NY, USA, 1153–1162. https://doi.org/10.1145/1124772.1124945

[50] Renwen Zhang, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, and Yi-Chieh Lee. 2024. The Dark Side of AI Companionship: A Taxonomy of Harmful Algorithmic Behaviors in Human-AI Relationships. https://doi.org/10.48550/arXiv.2410.20130 arXiv:2410.20130 [cs] version: 2.

[51] Tao Zhang, Chao Feng, Hui Chen, and Junjie Xian. 2022. Calming the customers by AI: Investigating the role of chatbot acting-cute strategies in soothing negative customer emotions. *Electronic Markets* 32, 4 (Dec. 2022), 2277–2292. https://doi.org/10.1007/s12525-022-00596-2 Num Pages: 2277-2292 Place: Heidelberg, Netherlands Publisher: Springer Nature B.V..